Exploiting Statistical and Relational Information on the Web and in Social Media

Lise Getoor & Lilyana Mihalkova
Department of Computer Science
University of Maryland
College Park, MD 20742

Abstract

This tutorial will provide an overview of statistical relational learning and inference techniques, motivating and illustrating them using web and social media applications. We will start by briefly surveying some of the sources of statistical and relational information on the web and in social media and will then dedicate most of the tutorial time to an introduction to representations and techniques for learning and reasoning with multi-relational information, viewing them through the lens of web and social media domains. We will end with a discussion of current trends and related fields, such as privacy in social networks.

1 Tutorial Overview

The growing popularity of Web 2.0, characterized by a proliferation of social media sites, and Web 3.0, with more richly semantically annotated objects and relationships, brings to light a variety of important prediction, ranking and extraction tasks. The input to these tasks is often best seen as some type of (noisy) multi-relational graph, which can be constructed in a variety of ways. For example, users' behaviors on the Web define a graph, the click graph, which can be viewed as a bipartite graph with queries and URLs, where the edges are weighted by the number of times a user clicked on a URL given a particular query. Social media sites often capture more semantically rich relationships such as friendship and affiliations. In addition to clicking, there are also many actions that users can take such as making a posting, asking a question, rating an item, purchasing an item and so on.

The common tasks that we want to perform in these situations typically depend both on the relational information contained in the graph, and also on statistics that are computed over the graph. In the first part of this tutorial, we will briefly survey several common Web applications, and show how they can be described as reasoning over multi-relational graphs. In the second part of the tutorial, we will describe techniques developed in the field of statistical relational learning (SRL). SRL addresses the challenge of learning from multi-relational data and representing the acquired knowledge in a form that allows for probabilistic inference. Two of the hallmarks of SRL approaches are, first, the ability to learn, represent, and leverage the rich sets of relations present among the entities, and, second, the ability to perform effective learning and reasoning in the presence of a considerable amount of noise and uncertainty. We argue in favor of using SRL techniques for Web problems by pointing out that these two main characteristics of SRL problems also typify problems that arise on the Web and discuss in detail several successful SRL applications on the Web.

2 Content

The tutorial web page can be found at
http://www.cs.umd.edu/projects/linqs/Tutorials/SRL-Web-SDM11/Home.html.

Part 1: Overview of Statistical and Relational Information on the Web. In the first part of the tutorial, we will provide a brief overview of some sources of relational information on the web.

Part 2: Toolkit. This is the main part of the tutorial. It will provide a survey of statistical relational learning and inference techniques, motivating and illustrating them from the point of view of web applications. This part will be structured as follows. First, we will discuss "flat" relational models in which relational information has been flattened by computing aggregates over relations, thus allowing training instances to be described by fixed-length feature vectors. We will give examples of web applications that follow this approach, e.g., [18]. Second, we will present collective models in which the labels of related instances are predicted collectively, thus allowing for assignments to be coordinated and some mistakes fixed. Here will discuss two collective approaches: the ICA algorithm [13, 11] and graphical models. The discussion of graphical models will lead us to the third subsection of part 2, in which we will present a survey of statistical relational learning (SRL). Here, we will overview some of the existing SRL representations, focusing on ones that either have been used on Web tasks, or are likely to scale to such tasks, such as probabilistic relational models [8], relational Markov networks [19], Markov logic networks [17], relational dependency networks [14], probabilistic similarity logic [5]. We will then describe learning and inference techniques for these models.

Part 3: Current Developments and Related Fields. In this final part we will discuss recent developments and future challenges. Topics here include lifted inference e.g., [16, 6] and privacy e.g., [21].

3 Acknowledgment

LM is supported by a CI Fellowship under NSF Grant #0937060 to the Computing Research Association.

References

[1] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. In *Proceedings of the 12th International World Wide Web Conference (WWW-03)*, 2003.

[2] R. Baeza-Yates and A. Tiberi. Extracting semantic relations from query logs. In *Proceedings of the 13th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD-07)*, 2007.

[3] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: Model and applications. In *Proceedings of the 17th Conference on Information and Knowledge Management (CIKM-07)*, 2008.

[4] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network analysis of collaboration structure in Wikipedia. In *Proceedings of the 18th International World Wide Web Conference (WWW-09)*, 2009.

[5] M. Broecheler, L. Mihalkova, and L. Getoor. Probabilistic similarity logic. In *Proceedings of 26th Conference on Uncertainty in Artificial Intelligence (UAI-10)*, 2010.

[6] R. de Salvo Braz, E. Amir, and D. Roth. Lifted first-order probabilistic inference. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, 2005.

[7] Z. Dou, R. Song, and J. Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th International World Wide Web Conference (WWW-07)*, 2007.

[8] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of relational structure. In *Proceedings of the 18th International Conference on Machine Learning (ICML-01)*, 2001.

[9] R. Guha, R. Kumar, P. Raghava, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th International World Wide Web Conference (WWW-04)*, 2004.

[10] R. Jones and K. L. Klinkner. Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th Conference on Information and Knowledge Management (CIKM-07)*, 2008.

[11] Q. Lu and L. Getoor. Link-based classification. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 2003.

[12] L. Mihalkova and R. J. Mooney. Learning to disambiguate search queries from short sessions. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD-09)*, 2009.

[13] J. Neville and D. Jensen. Iterative classification in relational data. In *Proceedings of the Workshop on Statistical Relational Learning at the 17th National Conference on Artificial Intelligence*, 2000.

[14] J. Neville and D. Jensen. Relational dependency networks. *Journal of Machine Learning Research*, 8:653–692, 2007.

[15] A. Plangprasopchok, K. Lerman, and L. Getoor. Growing a tree in the forest: constructing folksonomies by integrating structured metadata. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010.

[16] D. Poole. First-order probabilistic inference. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)*, 2003.

[17] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62:107–136, 2006.

[18] M. Richardson, E. Dominowska, and R. Ragno. Predicting clicks: Estimating the click-through rate for new ads. In *Proceedings of the 16th International World Wide Web Conference (WWW-07)*, 2007.

[19] B. Taskar, P. Abbeel, and D. Koller. Discriminative probabilistic models for relational data. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-02)*, 2002.

[20] D. Watts and P. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 34(4):441–558, 2007.

[21] E. Zheleva and L. Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th International World Wide Web Conference (WWW-09)*, 2009.